Wikipedia — Scientific Journal

vol. 0

January — April 2014

Wikipedia
— Scientific
Journal

1.

Citation:
O'Neill K, Aghaeepour N, Špidlen J, Brinkman R (2013) **Flow Cytometry Bioinformatics.** PLoS Comput Biol 9(12): e1003365. doi:10.1371/journal.pcbi.1003365

Flow Cytometry Bioinformatics

Autors:
Kieran O'Neill, *Co-lead authors*
Affiliations: Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada, Bioinformatics Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada.

Nima Aghaeepour, *Co-lead authors*
Affiliations: Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada, Bioinformatics Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada

Josef Špidlen
Affiliation: Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada

Ryan Brinkman, *rbrinkman@bccrc.ca*
Affiliations: Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

Abstract

Flow cytometry bioinformatics is the application of bioinformatics to flow cytometry data, which involves storing, retrieving, organizing, and analyzing flow cytometry data using extensive computational resources and tools. Flow cytometry bioinformatics requires extensive use of and contributes to the development of techniques from computational statistics and machine learning. Flow cytometry and related methods allow the quantification of multiple independent biomarkers on large numbers of single cells. The rapid growth in the multidimensionality and throughput of flow cytometry data, particularly in the 2000s, has led to the creation of a variety of computational analysis methods, data standards, and public databases for the sharing of results. Computational methods exist to assist in the preprocessing of flow cytometry data, identifying cell populations within it, matching those cell populations across samples, and performing diagnosis and discovery using the results of previous steps. For preprocessing, this includes compensating for spectral overlap, transforming data onto scales conducive to visualization and analysis, assessing data for quality, and normalizing data across samples and experiments. For population identification, tools are available to aid traditional manual identification of populations in

two-dimensional scatter plots (gating), to use dimensionality reduction to aid gating, and to find populations automatically in higher dimensional space in a variety of ways. It is also possible to characterize data in more comprehensive ways, such as the density-guided binary space partitioning technique known as probability binning, or by combinatorial gating. Finally, diagnosis using flow cytometry data can be aided by supervised learning techniques, and discovery of new cell types of biological importance by high-throughput statistical methods, as part of pipelines incorporating all of the aforementioned methods. Open standards, data, and software are also key parts of flow cytometry bioinformatics. Data standards include the widely adopted Flow Cytometry Standard (FCS) defining how data from cytometers should be stored, but also several new standards under development by the International Society for Advancement of Cytometry (ISAC) to aid in storing more detailed information about experimental design and analytical steps. Open data is slowly growing with the opening of the CytoBank database in 2010 and FlowRepository in 2012, both of which allow users to freely distribute their data, and the latter of which has been recommended as the preferred repository for MIFlowCyt-compliant data by ISAC. Open software is most widely available in the form of a suite of Bioconductor packages, but is also available for web execution on the GenePattern platform.

Flow Cytometry Data

Flow cytometers operate by hydrodynamically focusing suspended cells so that they separate from each other within a fluid stream. The stream is passed by one or more lasers, and the resulting fluorescent and scattered light is detected by photomultipliers. By using optical filters, particular fluorophores on or within the cells can be quantified by peaks in their emission spectra. This process is illustrated in Figure 1. Reporter molecules may be endogenous fluorophores such as chlorophyll or transgenic green fluorescent protein, or they may be artificial fluorophores covalently bonded to detection molecules such as antibodies for detecting proteins, or hybridization probes for detecting DNA or RNA.

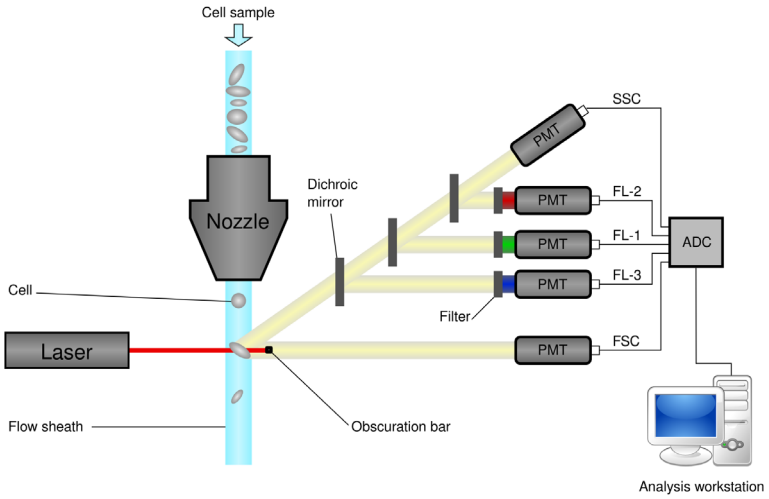

Figure 1.
Schematic diagram of a flow cytometer, showing focusing of the fluid sheath, laser, optics (in simplified form, omitting focusing), photomultiplier tubes (PMTs), analogue-to-digital converter, and analysis workstation.
doi:10.1371/journal.pcbi.1003365.g001

The ability to quantify these has led to flow cytometry being used in a wide range of applications, including but not limited to:

1. Monitoring of CD4 count in HIV [1]
2. Diagnosis of various cancers [2],[3]
3. Analysis of aquatic microbiomes [4]
4. Sperm sorting [5]
5. Measuring telomere length [6]

Until the early 2000s, flow cytometry could only measure a few fluorescent markers at a time. Through the late 1990s into the mid-2000s, however, rapid development of new fluorophores resulted in modern instruments capable of quantifying up to 18 markers per cell [7]. More recently, the new technology of mass cytometry replaces fluorophores with rare earth elements detected by time of flight mass spectrometry, achieving the ability to measure the expression of 34 or more markers [8]. At the same time, microfluidic qPCR methods are providing a flow cytometry–like method of quantifying 48 or more RNA molecules per cell [9]. The rapid increase in the dimensionality of flow cytometry data coupled with the development of high-throughput robotic platforms capable of assaying hundreds to thousands of samples automatically have created a need for improved computational analysis methods [7].

Steps in Computational Flow Cytometry Data Analysis

he process of moving from primary FCM data to disease diagnosis and biomarker discovery (illustrated in Figure 2) involves four major steps:

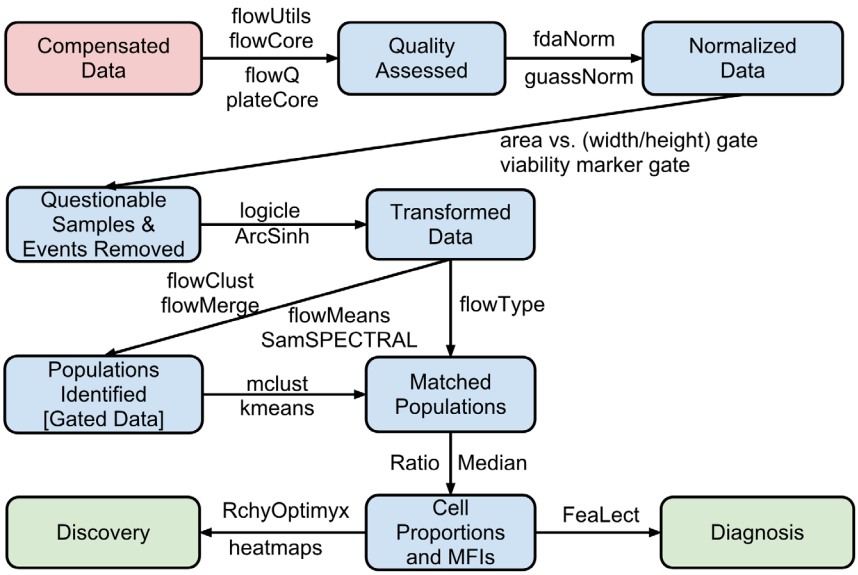

Figure 2.
An example pipeline for analysis of FCM data and some of the Bioconductor packages relevant to each step.
doi:10.1371/journal.pcbi.1003365.g002

1. Data preprocessing (including compensation, transformation, and normalization)
2. Cell population identification (a.k.a. gating)
3. Cell population matching for cross-sample comparison
4. Relating cell populations to external variables (diagnosis and discovery)

Saving of the steps taken in a particular flow cytometry workflow is supported by some flow cytometry software, and is important for the reproducibility of flow cytometry experiments. However, saved workspace files are rarely interchangeable between software [10]. An attempt to solve this problem is the development of the Gating-ML XML-based data standard (discussed in more detail in the data formats and interchange section), which is slowly being adopted in both commercial and open-source flow cytometry software [11].

Data Preprocessing

Compensation

When more than one fluorochrome is used with the same laser, their emission spectra frequently overlap. Each particular fluorochrome is typically measured using a bandpass optical filter set to a narrow band at or near the fluorochrome's emission intensity peak. The result is that the reading for any given fluorochrome is actually the sum of that fluorochrome's peak emission intensity, and the intensity of all other fluorochromes' spectra where they overlap with that frequency band. This overlap is termed spillover, and the process of removing spillover from flow cytometry data is called compensation [12].

Compensation is typically accomplished by running a series of representative samples each stained for only one fluorochrome, to give measurements of the contribution of each fluorochrome to each channel [12]. The total signal to remove from each channel can be computed by solving a system of linear equations based on this data to produce a spillover matrix, which when inverted and multiplied with the raw data from the cytometer produces the compensated data [12], [13]. The processes of computing the spillover matrix, or applying a precomputed spillover matrix to compensate flow cytometry data, are standard features of flow cytometry software [14].

Quality Control

Particularly in newer, high-throughput experiments, there is a need for visualization methods to help detect technical errors in individual samples. One approach is to visualize summary statistics, such as the empirical distribution functions of single dimensions of technical or biological replicates to ensure they are the similar [22]. For more rigor, the KolmogorovSmirnov test can be used to determine if individual samples deviate from the norm [22]. The Grubbs test for outliers may be used to detect samples deviating from the group.

A method for quality control in higher-dimensional space is to use probability binning with bins fit to the whole dataset pooled together [23]. Then the standard deviation of the number of cells falling in the bins within each sample can be taken as a measure of multidimensional similarity, with samples that are closer to the norm having a smaller standard deviation [23]. With this method, higher standard deviation can indicate outliers, although this is a relative measure as the absolute value depends partly on the number of bins.

With all of these methods, the cross-sample variation is being measured. However, this is the combination of technical variations introduced by the instruments and handling, and actual biological information that is desired to be measured. Disambiguating the technical and the biological contributions to between-sample variation can be a difficult to impossible task [24].

Normalization

Particularly in multicenter studies, technical variation can make biologically equivalent populations of cells difficult to match across samples. Normalization methods to remove technical variance, frequently derived from image registration techniques, are thus a critical step in many flow cytometry analyses. Single-marker normalization can be performed using landmark registration, in which peaks in a kernel density estimate of each sample are identified and aligned across samples [24].

Identifying Cell Populations

A critical step in analysis of flow cytometric data is the identification of multidimensional regions that contain functionally and phenotypically homogeneous groups of cells for further analysis [27].

Gating

The data generated by flow cytometers can be plotted in one or two dimensions to produce a histogram or scatter plot. The regions on these plots can be sequentially separated, based on fluorescence intensity, by creating a series of subset extractions, termed "gates." These gates can be produced using software, e.g., FlowJo [28], FCS Express [29], WinMDI [30], CytoPaint (aka Paint-A-Gate) [31], VenturiOne, CellQuest Pro, Cytospec [32], or Kaluza [33].

In datasets with a low number of dimensions and limited cross-sample technical and biological variability (e.g., clinical laboratories), manual analysis of specific cell populations can produce effective and reproducible results. However, exploratory analysis of a large number of cell populations in a high-dimensional dataset is not feasible [34]. In addition, manual analysis in less controlled settings (e.g., cross-laboratory studies) can increase the overall error rate of the study [35]. In one study, several computational gating algorithms performed better than manual analysis in the presence of some variation [26] (illustrated in Figure 3). However, despite the considerable advances in computational analysis, manual gating remains the main solution for the identification of specific rare cell populations that are not well-separated from other cell types.

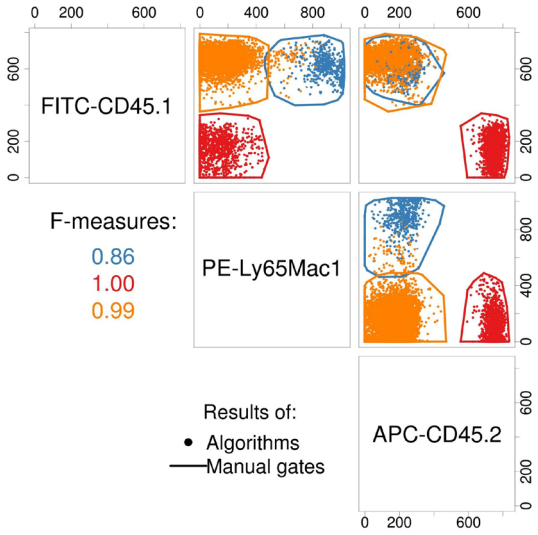

Figure 3.
Comparison of consensus of eight independent manual gates (polygons) and automated gates (colored dots).
The consensus of the manual gates and the algorithms were produced using the CLUE package [25]. Figure reproduced with permission from [26]. doi:10.1371/journal.pcbi.1003365.g003

Gating guided by dimension reduction.

As the number of markers measured by flow cytometry increases, the number of scatter plots that need to be investigated increases exponentially (some markers need to be investigated several times for each group of cells to resolve high-dimensional differences between cell types that appear to be similar in most markers) [36]. To address this issue, principal component analysis has been used to summarize the high-dimensional datasets using a combination of markers that maximizes the variance of all data points [37]. However, PCA is a linear method and is not able to preserve complex and non-linear relationships. More recently, two-dimensional minimum spanning tree layouts have been used to guide the manual gating process (illustrated in Figure 4). Density-based down-sampling and clustering was used to better represent rare populations and control the time and memory complexity of the minimum spanning tree construction process [38]. More sophisticated dimension reduction algorithms are yet to be investigated [39].

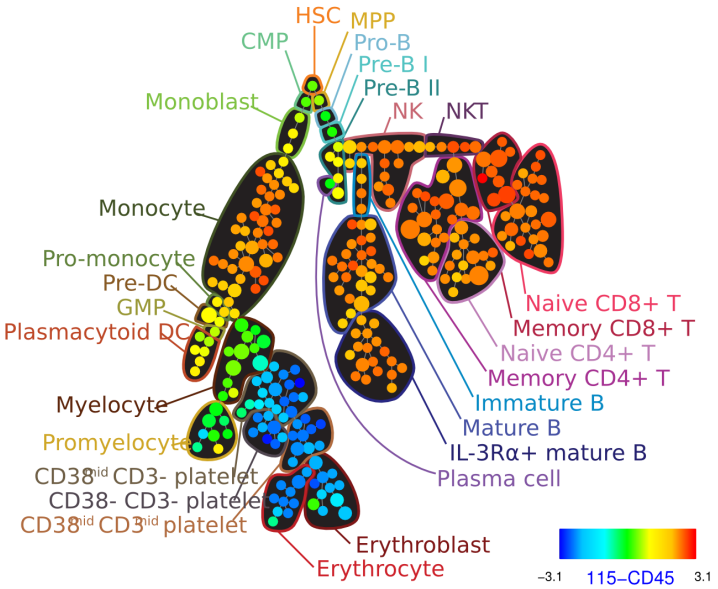

Figure 4.
Cell populations in a high-dimensional mass-cytometry dataset manually gated after dimension reduction using 2-D layout for a minimum spanning tree. Figure reproduced from the data provided in [40].
doi:10.1371/journal.pcbi.1003365.g004

Automated gating

Developing computational tools for identification of cell populations has been an area of active research only since 2008. Many individual clustering approaches have recently been developed, including model-based algorithms (e.g., flowClust [41] and FLAME [42]), density-based algorithms (e.g., FLOCK [43] and SWIFT), graph-based approaches (e.g., SamSPECTRAL [44]), and, most recently, hybrids of several approaches (flowMeans [45] and flowPeaks [46]). These algorithms are different in terms of memory and time complexity, their software requirements, their ability to automatically determine the required number of cell populations, and their sensitivity and specificity. The FlowCAP (Flow Cytometry: Critical Assessment of Population Identification Methods) project, with active participation from most academic groups with research efforts in the area, is providing a way to objectively cross-compare state-of-the-art automated analysis approaches [26].

Probability Binning Methods

Probability binning is a non-gating analysis method in which flow cytometry data is split into quantiles on a univariate basis (shown in Figure 5) [47]. The locations of the quantiles can then be used to test for differences between samples using the chi-squared test [47].

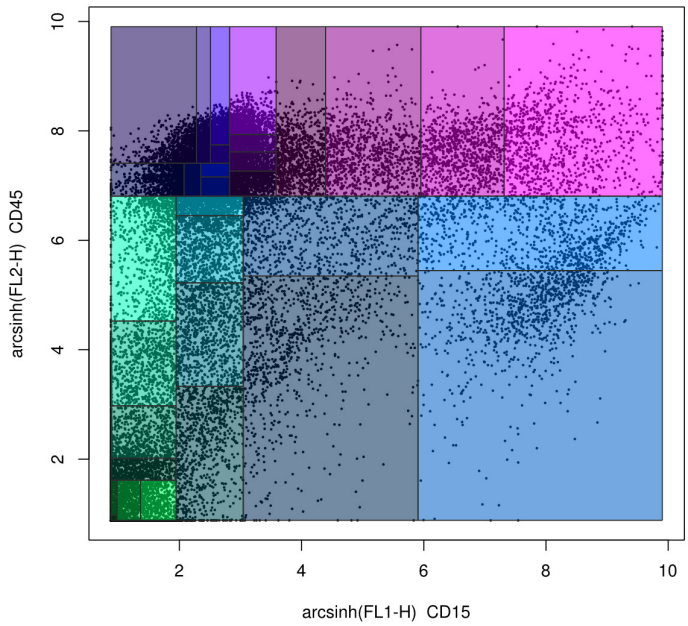

Figure 5.
An example of probability binning, created using the flowFP Bioconductor package. The dots represent individual events in an FCS file. The rectangles represent the bins. doi:10.1371/journal.pcbi.1003365.g005

This was later extended into multiple dimensions in the form of frequency difference gating, a binary space partitioning technique in which data is iteratively partitioned along the median [48]. These partitions (or bins) are fit to a control sample. Then the proportion of cells falling within each bin in test samples can be compared to the control sample by the chi-squared test.

Finally, cytometric fingerprinting uses a variant of frequency difference gating to set bins and measure for a series of samples how many cells fall within each bin [23]. These bins can be used as gates and used for subsequent analysis similarly to automated gating methods.

Combinatorial Gating

High-dimensional clustering algorithms are often unable to identify rare cell types that are not well separated from other major populations. Matching these small cell populations across multiple samples is even more challenging. In manual analysis, prior biological knowledge (e.g., biological controls) provides guidance to reasonably identify these populations. However, integrating this information into the exploratory clustering process (e.g., as in semi-supervised learning) has not been successful.

An alternative to high-dimensional clustering is to identify cell populations using one marker at a time and then combine them to produce higher-dimensional clusters. This functionality was first implemented in FlowJo [28]. The flowType algorithm builds on this framework by allowing the exclusion of the markers [49]. This enables the development of statistical tools (e.g., RchyOptimyx) that can investigate the importance of each marker and exclude high-dimensional redundancies [50].

Diagnosis and Discovery

After identification of the cell population of interest, a cross-sample analysis can be performed to identify phenotypical or functional variations that are correlated with an external variable (e.g., a clinical outcome). These studies can be partitioned into two main groups:

Diagnosis

In these studies, the goal usually is to diagnose a disease (or a sub-class of a disease) using variations in one or more cell populations. For example, one can use multidimensional clustering to identify a set of clusters, match them across all samples, and then use supervised learning to construct a classifier for prediction of the classes of interest (e.g., this approach can be used to improve the accuracy of the classification of specific lymphoma subtypes [51]). Alternatively, all the cells from the entire cohort can be pooled into a single multidimensional space for clustering before classification [52]. This approach is particularly suitable for datasets with a high amount of biological variation (in which cross-sample matching is challenging) but requires technical variations to be carefully controlled [53].

Discovery

In a discovery setting, the goal is to identify and describe cell populations correlated with an external variable (as opposed to the diagnosis setting in which the goal is to combine the predictive power of multiple cell types to maximize the accuracy of the results). Similar to the diagnosis use case, cluster matching in high-dimensional space can be used for exploratory analysis, but the descriptive power of this approach is very limited, as it is hard to characterize and visualize a cell population in a high-dimensional space without first reducing the dimensionality [52], [54]. Finally, combinatorial gating approaches have been particularly successful in exploratory analysis of FCM data. Simplified Presentation of Incredibly Complex Evaluations (SPICE) is a software package that can use the gating functionality of FlowJo to statistically evaluate a wide range of different cell populations and visualize those that are correlated with the external outcome. flowType and RchyOptimyx (as previously discussed) expand this technique by adding the ability of exploring the impact of independent markers on the overall correlation with the external outcome. This enables the removal of unnecessary markers and provides a simple visualization of all identified cell types. In a recent analysis of a large (n = 466) cohort of HIV+ patients, this pipeline identified three correlates of protection against HIV, only one of which had been previously identified through extensive manual analysis of the same dataset (as illustrated in Figure 6) [49].

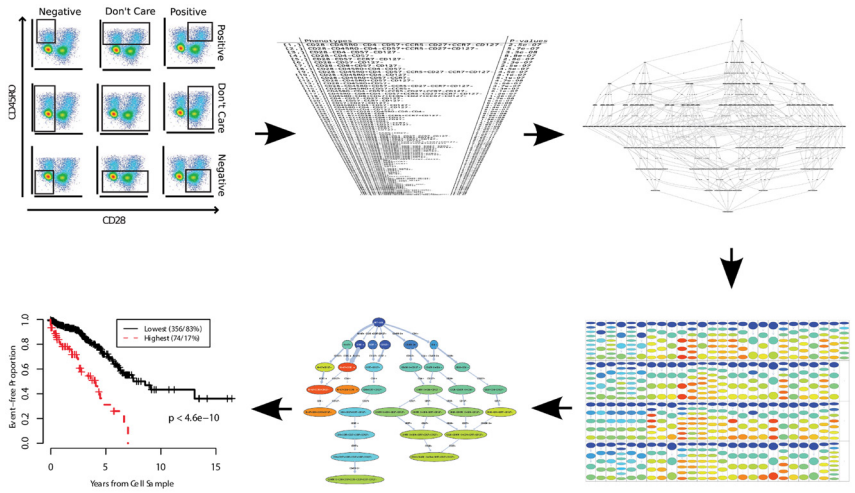

Figure 6.
Overview of the flowType/RchyOptimyx pipeline for identification of correlates of protection against HIV.
First, tens of thousands of cell populations are identified by combining one-dimensional partitions (panel 1). The cell populations are then analyzed using a statistical test (and Bonferroni's method for multiple testing correction) to identify those correlated with the survival information. Panel 3 shows a complete gating hierarchy describing all possible strategies for gating that cell population. This graph can be mined to identify the "best" gating strategy (i.e., the one in which the most important markers appear earlier). These hierarchies for all selected phenotypes are demonstrated in panel 4. In panel 5, these hierarchies are merged into a single graph that summarizes the entire dataset and demonstrates the trade-off between the number of markers involved in each phenotype and the significance of the correlation with the clinical outcome (e.g., as measured by the KaplanMeier estimator in panel 6). Figure reproduced in part from [49] (public domain) and [50].
doi:10.1371/journal.pcbi.1003365.g006

Data Formats and Interchange

Flow Cytometry Standard

Flow cytometry data is typically saved for analysis in the form of an array, with fluorescence and scatter channels represented in columns and individual "events" (most of which are cells) forming the rows, as shown in Figure 7. The number of events acquired from each sample usually ranges between the low thousands and the low millions.

Markers

Cells

	FSC-A	FSC-H	SSC-A	B515-A	R780-A	R710-A	R660-A	V800-A	V655-A	V585-A	V450-A	G780-A	G710-A	G660-A	G610-A	G560-A
[1,]	27700.75	27291.75	177.52585	1984.485	625.0796	1232.1008	748.5101	1553.0295	1350.2565	3175.717	2338.1038	2286.1709	1758.4244	2550.914	1862.9843	1972.4854
[2,]	41264.25	39764.25	320.12296	3639.620	539.7032	1433.3112	1470.2659	2217.6750	2305.3516	5683.995	4767.4976	2617.9441	2063.0974	3401.507	2471.2236	2268.8318
[3,]	65054.75	57606.25	203.01607	2191.861	198.6541	726.9798	766.2198	802.2521	809.9579	1763.534	2870.2039	721.3581	750.2025	1156.660	879.6395	802.7821
[4,]	30584.00	31664.50	130.68690	1873.409	1304.0895	2528.7083	784.6980	1702.3671	1185.8608	2063.256	1839.7687	3945.7107	2480.2170	2585.438	1850.5399	2768.8843
[5,]	39505.75	39626.00	203.25166	2540.620	323.2625	857.1525	715.0004	1117.4775	1746.5798	3810.514	3442.1965	1766.7188	1595.8044	2746.546	2118.4902	1560.5364
[6,]	33171.50	34794.00	333.64246	2192.864	1408.8563	2573.5095	1604.2236	2128.1748	1727.5891	3734.910	2551.4509	5498.1421	2989.3887	3295.179	2938.3291	4663.7070
[7,]	63711.00	54475.50	1122.48340	3879.044	1730.8085	3573.5652	1691.8744	5106.0596	3578.0332	9183.305	7264.9624	7268.8550	4263.6313	4767.073	4025.7312	5758.2100
[8,]	40000.75	40213.50	236.54262	2545.858	1081.6753	2313.5962	1411.0983	2989.7524	1920.4047	4386.833	3081.6401	3615.0461	2552.7622	3304.552	2666.3101	3532.8064
[9,]	49286.00	49182.50	78.61845	1601.092	123.2834	493.6364	242.0255	633.3533	759.2227	1920.717	2082.5884	996.4465	805.6166	1197.744	957.8304	937.6376
[10,]	32209.75	33368.25	203.29897	2387.361	1056.0723	1769.4005	939.7758	1693.8635	1579.7000	3308.004	2106.1416	2835.7053	2057.3567	2653.473	1879.0623	2306.1431
[11,]	35937.25	36212.50	220.66580	2901.591	1218.1395	3202.3853	1059.7604	2443.0205	2253.0146	4684.219	2864.9727	5582.5425	3703.8475	4023.313	3111.4429	4476.1655
[12,]	32905.50	33897.50	233.98033	2726.240	1952.0721	3405.7139	2726.1091	2988.6882	2011.0159	4644.849	2661.8582	7080.1904	3540.6592	3915.066	3389.4014	6423.2324
[13,]	36028.50	35845.50	219.18674	3221.668	2542.3389	3895.0371	2283.0444	3331.8298	2479.6580	5253.382	2692.4919	9087.0859	4306.8105	4652.118	4095.7131	7948.1416
[14,]	38616.00	38775.00	218.46669	3218.305	582.6801	1022.7971	1255.5858	2150.4185	1993.2681	4657.159	1881.8761	2058.4124	1753.6392	2576.016	1914.5718	2386.2805
[15,]	45282.25	42223.25	1173.74487	6941.545	705.4651	1649.9570	1615.0811	4287.2036	3778.2302	11158.718	8247.2236	3138.5728	2382.1287	3487.720	2436.5886	3280.1677
[16,]	36246.25	36207.75	189.15569	3049.417	1736.7826	2823.7266	1031.0308	2824.6582	2053.6843	4689.628	2930.9685	6049.8901	3479.7673	3642.755	2885.2795	4767.1309
[17,]	29282.75	29884.00	209.64102	1836.197	612.2673	1149.7164	870.3303	1720.2170	1525.6914	2782.762	2335.3994	2180.0403	1825.0010	2702.035	1921.6849	1730.4067
[18,]	57757.25	54448.25	1999.17517	12972.877	4364.5908	11298.7070	6745.5039	20934.3457	17057.1934	30293.186	7082.0664	14765.4785	13275.3828	18369.047	14707.2549	9409.2637
[19,]	33301.00	33093.50	208.47151	2146.622	429.5022	855.5981	845.9418	1207.8969	1297.2683	2751.205	1081.5209	1716.1721	1479.2837	2110.357	1566.1290	1371.2155
[20,]	34478.25	35390.75	211.26921	3060.585	2016.3651	3442.5408	1348.4852	2673.9729	2259.8494	3781.046	1290.9741	7142.1543	3906.1033	4101.971	3586.8269	5697.7954
[21,]	29406.25	28219.50	231.55798	3008.380	997.8875	2319.5779	1514.2091	1757.2463	1675.9983	3019.569	1565.1250	4008.3364	2470.9639	2938.669	2289.2273	3631.8906
[22,]	49978.50	48517.75	537.04224	3122.343	981.1232	2252.1189	1861.3472	2518.4731	2230.5327	5010.112	3615.4170	4111.9141	2611.4641	3379.042	2634.6746	3722.1040
[23,]	39872.50	37620.75	198.75706	2719.222	1657.0939	2945.5713	1025.1293	2203.0527	1670.1367	3072.683	1879.3065	6735.8999	3531.4177	3581.812	2901.9954	4790.6758
[24,]	33395.00	35331.75	220.46056	2664.632	690.1926	1483.0898	1736.9537	1397.0316	1982.9124	3473.224	1987.4087	2142.7542	1793.1940	2844.251	2057.5803	2655.7776
[25,]	46976.00	47355.25	231.33037	2530.461	537.1376	1194.0681	1072.7083	1531.7494	1766.5841	3532.332	3953.0066	3967.8032	1918.4053	3249.490	2566.3657	3888.4016
[26,]	56663.75	51458.25	223.06416	3217.866	398.6222	1279.4880	1207.4561	1268.9905	1553.6884	3397.891	2623.9312	1686.8226	1583.9100	2461.450	1874.0762	2013.9438
[27,]	50818.75	48556.25	305.77182	3714.351	577.0732	1364.4095	1064.0983	1633.2513	2077.0466	3332.829	2654.2678	2074.0388	1966.8965	2876.737	1987.8170	2010.1973
[28,]	36225.25	36196.75	180.30524	2636.466	946.7570	2138.4143	1695.0502	1807.8429	2057.7292	3673.466	2611.5857	3530.6289	2630.3755	3608.507	2727.3037	2884.0103
[29,]	28509.25	30715.50	230.27397	1072.201	1867.2009	1643.1423	882.4811	1201.5806	688.1475	1810.280	616.8154	3604.4263	1530.6011	1353.209	1167.5325	2577.3455
[30,]	37198.75	36200.50	237.67776	3046.719	1376.3452	2580.9287	1326.2197	2599.6101	2196.7258	5525.008	3665.0623	4946.4224	3421.0840	3996.508	3045.4102	4413.0908

Figure 7.
Representation of flow cytometry data from an instrument with three scatter channels and 13 fluorescent channels.
Only the values for the first 30 (of hundreds of thousands) of cells are shown.
doi:10.1371/journal.pcbi.1003365.g007

The first version of a Flow Cytometry Standard (FCS) was developed in 1984 [55]. Since then, FCS has become the standard file format supported by all flow cytometry software and hardware vendors. FCS is a binary file format with three main segments: a text segment containing metadata in keyword/value pairs structures, a data segment usually containing a matrix of detected expression values (so-called list mode format), and a rarely used analysis segment. The FCS specification has traditionally been developed and maintained by the International Society for Advancement of Cytometry (ISAC) [56].

Over the years, updates have been incorporated to adapt to technological advancements in both flow cytometry and computing technologies.

In 1990, FCS 2.0 was introduced [57]. Features introduced in FCS 2.0 included the option of multiple datasets within a data file, the use of different byte orders accommodating hardware variations on different computing platforms, and basic compensation and scaling information. FCS 2.0 was followed by FCS 3.0 in 1997, which introduced the possibility of storing datasets larger than 100 MB [58].

The latest version, FCS 3.1, was introduced in 2010 [59]. It retains the basic FCS file structure and most features of previous versions of the standard. Changes included in FCS 3.1 address potential ambiguities in the previous versions and provide a more robust standard. They include simplified support for international characters and improved support for storing compensation. The major additions are support for preferred display scale, a standardized way of capturing the sample volume, information about the origins of the data file, and support for plate and well identification in high-throughput, plate-based experiments.

FCS used to be the only widely adopted file format in flow cytometry. Recently, additional standard file formats have been developed by ISAC.

netCDF

ISAC is considering replacing FCS with a flow cytometry–specific version of the Network Common Data Form (netCDF) file format [60]. netCDF is a set of freely available software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. In 2008, ISAC drafted the first version of netCDF conventions for storage of raw flow cytometry data [61].

Archival Cytometry Standard (ACS)

The Archival Cytometry Standard (ACS) is being developed to bundle data with different components describing cytometry experiments [62]. It captures relations among data, metadata, analysis files, and other components, and includes support for audit trails, versioning, and digital signatures. The ACS container is based on the ZIP file format with an XML-based table of contents specifying relations among files in the container. The XML Signature W3C Recommendation has been adopted to allow for digital signatures of components within the ACS container. An initial draft of ACS was designed in 2007 and finalized in 2010. Since then, ACS support has been introduced in several software tools including FlowJo and Cytobank.

Gating-ML

The lack of gating interoperability has traditionally been a bottleneck preventing reproducibility of flow cytometry data analysis and the usage of multiple analytical tools. To address this shortcoming, ISAC developed Gating-ML, an XML-based mechanism to formally describe gates and related data (scale) transformations [10]. The draft recommendation version of Gating-ML was approved by ISAC in 2008, and it is partially supported by tools like FlowJo, the flowUtils library in R/Bioconductor, and FlowRepository [62]. It supports rectangular gates, polygon gates, convex polytopes, ellipsoids, decision trees, and Boolean collections of any of the other types of gates. In addition, it includes dozens of built-in public transformations that have been shown to be potentially useful for display or analysis of cytometry data. In 2013, Gating-ML version 2.0 was approved by ISAC's Data Standards Task Force as a Recommendation. This new version offers slightly less flexibility in terms of the power of gating description; however, it is also significantly easier to implement in software tools [11].

Classification Results (CLR)

The Classification Results (CLR) File Format [63] has been developed to exchange the results of manual gating and algorithmic classification approaches in a standard way in order to be able to report and process the classification. CLR is based on the commonly supported CSV file format with columns corresponding to different classes and cell values

containing the probability of an event being a member of a particular class. These are captured as values between 0 and 1. Simplicity of the format and its compatibility with common spreadsheet tools have been the major requirements driving the design of the specification. Although it was originally designed for the field of flow cytometry, it is applicable in any domain that needs to capture either fuzzy or unambiguous classifications of virtually any kinds of objects.

Public Data and Software

As in other bioinformatics fields, development of new methods has primarily taken the form of free open-source software, and several databases have been created for depositing open data.

Bioconductor

The Bioconductor project is a repository of free open-source software, mostly written in the R programming language [64]. As of July 2013, Bioconductor contained 21 software packages for processing flow cytometry data [65]. These packages cover most of the range of functionality described earlier in this article.

GenePattern

GenePattern is a predominantly genomic analysis platform with over 200 tools for analysis of gene expression, proteomics, and other data. A web-based interface provides easy access to these tools and allows the creation of automated analysis pipelines enabling reproducible research. Recently, a GenePattern Flow Cytometry Suite has been developed in order to bring advanced flow cytometry data analysis tools to experimentalists without programmatic skills. It contains close to 40 open-source GenePattern flow cytometry modules covering methods from basic processing of flow cytometry standard (i.e., FCS) files to advanced algorithms for automated identification of cell populations, normalization, and quality assessment. Internally, most of these modules leverage from functionality developed in Bioconductor.

Much of the functionality of the Bioconductor packages for flow cytometry analysis has been packaged up for use with the GenePattern [66] workflow system, in the form of the GenePattern Flow Cytometry Suite [67].

Public Databases

The Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) requires that any flow cytometry data used in a publication be available, although this does not include a requirement that it be deposited in a public database [68]. Thus, although the journals Cytometry A and B, as well as all journals from the Nature Publishing Group require MIFlowCyt compliance, there is still relatively little publicly available flow cytometry data. Some efforts have been made toward creating public databases, however.

Firstly, CytoBank, which is a complete web-based flow cytometry data storage and analysis platform, has been made available to the public in a limited form [69]. Using the CytoBank code base, FlowRepository was developed in 2012 with the support of ISAC to be a public repository of flow cytometry data [70]. FlowRepository facilitates MIFlowCyt compliance [71], and as of July 2013 contained 65 public datasets [72].

Datasets

In 2012, the flow cytometry community started to release a set of publicly available datasets. A subset of these datasets representing the existing data analysis challenges is described below. For comparison against manual gating, the FlowCAP-I project has released five datasets, manually gated by human analysts, and two of them gated by eight independent analysts [26]. The FlowCAP-II project included three datasets for binary classification and also reported several algorithms that were able to classify these samples perfectly. FlowCAP-III included two larger datasets for comparison against manual gates as well as one more challenging sample classification dataset. As of March 2013, public release of FlowCAP-III was still in progress [73]. The datasets used in FlowCAP-I, II, and III either have a low number of subjects or parameters. However, recently several more complex clinical datasets have been released including a dataset of 466 HIV-infected subjects, which provides both 14 parameter assays and sufficient clinical information for survival analysis [50], [74]–[76].

Another class of datasets are higher-dimensional mass cytometry assays. A representative of this class of datasets is a study that includes analysis of two bone marrow samples using more than 30 surface or intracellular markers under a wide range of different stimulations [8]. The raw data for this dataset is publicly available as described in the manuscript, and manual analyses of the surface markers are available upon request from the authors.

Open Problems

Despite rapid development in the field of flow cytometry bioinformatics, several problems remain to be addressed.

Variability across flow cytometry experiments arises from biological variation among samples, technical variations across instruments used, as well as methods of analysis. In 2010, a group of researchers from Stanford University and the National Institutes of Health pointed out that while technical variation can be ameliorated by standardizing sample handling, instrument setup, and choice of reagents, solving variation in analysis methods will require similar standardization and computational automation of gating methods [77]. They further opined that centralization of both data and analysis could aid in decreasing variability between experiments and in comparing results [77].

This was echoed by another group of Pacific Biosciences and Stanford University researchers, who suggested that cloud computing could enable centralized, standardized, high-throughput analysis of flow cytometry experiments [78]. They also emphasized that ongoing development and adoption of standard data formats could continue to aid in reducing variability across experiments [78]. They also proposed that new methods will be needed to model and summarize results of high-throughput analysis in ways that can be interpreted by biologists [78], as well as ways of integrating large-scale flow cytometry data with other high-throughput biological information, such as gene expression, genetic variation, metabolite levels, and disease states [78].

Acknowledgments

The version history of the text file and the peer reviews (and response to reviews) are available as supporting information in Text S1 and Text S2.

References

1. Brando B, Barnett D, Janossy G, Mandy F, Autran B, et al. (2000) Cytofluorometric methods for assessing absolute numbers of cell subsets in blood. European Working Group on Clinical Cell Analysis. Cytometry 42: 327–346. doi: 10.1002/1097-0320(20001215)42:6<327::aid-cyto1000>3.0.co;2-f

2. Ferreira-Facio CS, Milito C, Botafogo V, Fontana M, Thiago LS, et al. (2013) Contribution of multiparameter flow cytometry immunophenotyping to the diagnostic screening and classification of pediatric cancer. PLoS ONE 8: e55534 doi:10.1371/journal.pone.0055534.

3. Wu D, Wood BL, Fromm JR (2013) Flow cytometry for non-Hodgkin and classical Hodgkin lymphoma. Methods Mol Biol 971: 27–47. doi: 10.1007/978-1-62703-269-8_2

4 Wang Y, Hammes F, De Roy K, Verstraete W, Boon N (2010) Past, present and future applications of flow cytometry in aquatic microbiology. Trends Biotechnol 28: 416–424. doi: 10.1016/j.tibtech.2010.04.006

5. Johnson LA, Flook JP, Look MV, Pinkel D (1987) Flow sorting of X and Y chromosome-bearing spermatozoa into two populations. Gamete Res 16: 1–9. doi: 10.1002/mrd.1120160102

6. Baerlocher GM, Vulto I, de Jong G, Lansdorp PM (2006) Flow cytometry and FISH to measure the average length of telomeres (flow FISH). Nat Protoc 1: 2365–2376. doi: 10.1038/nprot.2006.263

7. Chattopadhyay PK, Hogerkorp C-M, Roederer M (2008) A chromatic explosion: the development and future of multiparameter flow cytometry. Immunology 125: 441–449. doi: 10.1111/j.1365-2567.2008.02989.x

8. Behbehani GK, Bendall SC, Clutter MR, Fantl WJ, Nolan GP (2012) Single-cell mass cytometry adapted to measurements of the cell cycle. Cytometry A 81: 552–566. doi: 10.1002/cyto.a.22075

9. White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, et al. (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci U S A 108: 13999–14004. doi: 10.1073/pnas.1019446108

10. Spidlen J, Leif RC, Moore W, Roederer M (2008) International Society for the Advancement of Cytometry Data Standards Task Force, (2008) et al. Gating-ML: XML-based gating descriptions in flow cytometry. Cytometry A 73A: 1151–1157. doi: 10.1002/cyto.a.20637

11. International Society for Advancement of Cytometry (2013) Gating-ML 2.0 (Report). Available: http://flowcyt.sourceforge.net/gating/latest.pdf. Accessed 8 November 2013.

12. Roederer M (2002) Compensation in flow cytometry. Curr Protoc Cytom Chapter 1: Unit 1.14.

13. Bagwell CB, Adams EG (1993) Fluorescence spectral overlap compensation for any number of flow cytometry parameters. Ann N Y Acad Sci 677: 167–184. doi: 10.1111/j.1749-6632.1993.tb38775.x

14. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, et al. (2009) flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics 10: 106. doi: 10.1186/1471-2105-10-106

15. Shapiro HM (2003) Practical flow cytometry. New York: Wiley-Liss. pp. 235.

16. Parks DR, Roederer M, Moore WA (2006) A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry A 69: 541–551. doi: 10.1002/cyto.a.20258

17. Finak G, Perez J-M, Weng A, Gottardo R (2010) Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics 11: 546. doi: 10.1186/1471-2105-11-546

18. Moore WA, Parks DR (2012) Update for the logicle data scale including operational code implementations. Cytometry A 81: 273–277. doi: 10.1002/cyto.a.22030

19 Bagwell CB (2005) Hyperlog-a flexible log-like transform for negative, zero, and positive valued data. Cytometry A 64: 34–42. doi: 10.1002/cyto.a.20114

20. Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73: 321–332. doi: 10.1002/cyto.a.20531

21. Qian Y, Liu Y, Campbell J, Thomson E, Kong YM, et al. (2012) FCSTrans: an open source software system for FCS file conversion and data transformation. Cytometry A 81: 353–356. doi: 10.1002/cyto.a.22037

22. Le Meur N, Rossini A, Gasparetto M, Smith C, Brinkman RR, et al. (2007) Data quality assessment of ungated flow cytometry data in high throughput experiments. Cytometry A 71: 393–403. doi: 10.1002/cyto.a.20396

23. Rogers WT, Moser AR, Holyst HA, Bantly A, Mohler ER, et al. (2008) Cytometric fingerprinting: quantitative characterization of multivariate distributions. Cytometry A 73: 430–441. doi: 10.1002/cyto.a.20545

24. Hahne F, Khodabakhshi AH, Bashashati A, Wong C-J, Gascoyne RD, et al. (2010) Per-channel basis normalization methods for flow cytometry data. Cytometry A 77: 121–131. doi: 10.1002/cyto.a.20823

25. Hornik K, Böhm W (2013) clue: Cluster ensembles, version 0.3-47. Available: http://cran.r-project.org/web/packages/clue/index.html. Accessed 15 February 2013.

26. Aghaeepour N, Finak G (2013) FlowCAP Consortium DREAM Consortium (2013) Hoos H, et al. (2013) Critical assessment of automated flow cytometry data analysis techniques. Nat Methods 10: 228–238. doi: 10.1038/nmeth0513-445c

27. Lugli E, Roederer M, Cossarizza A (2010) Data analysis in flow cytometry: the future just started. Cytometry A 77: 705–713. doi: 10.1002/cyto.a.20901

28. Tree Star, Inc (2013) FlowJo. Available: http://www.flowjo.com/vX/en/ws.boolean.html. Accessed 5 April 2013.

29. De Novo Software (2013) FCS Express. Available: http://www.denovosoftware.com/site/Flow-RUO-Overview.shtml. Accessed 3 April 2013.

30. Saluk A (2013) TSRI Cytometry Software Page. Available: http://facs.scripps.edu/software.html. Accessed September 3 2009.

31. Leukobyte (2013) CytoPaint Classic. Available: http://leukobyte.com Accessed 5 April 2013.

32. Purdue University Cytometry Laboratories (2013) PUCL Cytometry Software Page. Available: http://www.cyto.purdue.edu/Purdue_software. Accessed 7 July 2011.

33. Beckman Coulter (2013) Kaluza. Available: http://www.beckmancoulter.com. Accessed 10 February 2013.

34. Bendall SC, Nolan GP (2012) From single cells to deep phenotypes in cancer. Nat Biotechnol 30: 639–647. doi: 10.1038/nbt.2283

35. Maecker HT, Rinfret A, D'Souza P, Darden J, Roig E, et al. (2005) Standardization of cytokine flow cytometry assays. BMC Immunol 6: 13. doi: 10.1016/j.clim.2007.03.345

36. Virgo PF, Gibbs GJ (2012) Flow cytometry in clinical pathology. Ann Clin Biochem 49: 17–28. doi: 10.1258/acb.2011.011128

37. Costa ES, Pedreira CE, Barrena S, Lecrevisse Q, Flores J, et al. (2010) Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. Leukemia 24: 1927–1933. doi: 10.1038/leu.2010.160

38. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, et al. (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol 29: 886–891. doi: 10.1038/nbt.1991

39. van der Maaten L (2013) Matlab Toolbox for Dimensionality Reduction. Available: http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html. Accessed 10 February 2013.

40. Bendall SC, Simonds EF, Qiu P, Amir E-aD, Krutzik PO, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332: 687–696. doi: 10.1126/science.1198704

41. Lo K, Hahne F, Brinkman RR, Gottardo R (2009) flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10: 145. doi: 10.1186/1471-2105-10-145

42. Pyne S, Hu X, Wang K, Rossin E, Lin T-I, et al. (2009) Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci U S A 106: 8519–8524. doi: 10.1073/pnas.0903028106

43. Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, et al. (2010) Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry B Clin Cytom 78 Suppl 1: S69–82. doi: 10.1002/cyto.b.20554

44. Zare H, Shooshtari P, Gupta A, Brinkman RR (2010) Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinformatics 11: 403. doi: 10.1186/1471-2105-11-403

45. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR (2011) Rapid cell population identification in flow cytometry data. Cytometry A 79: 6–13. doi: 10.1002/cyto.a.21007

46. Ge Y, Sealfon SC (2012) flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics 28: 2052–2058. doi: 10.1093/bioinformatics/bts300

47. Roederer M, Treister A, Moore W, Herzenberg LA (2001) Probability binning comparison: a metric for quantitating univariate distribution differences. Cytometry 45: 37–46. doi: 10.1002/1097-0320(20010901)45:1<37::aid-cyto1142>3.0.co;2-e

48. Roederer M, Hardy RR (2001) Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry 45: 56–64. doi: 10.1002/1097-0320(20010901)45:1<56::aid-cyto1144>3.0.co;2-9

49. Aghaeepour N, Chattopadhyay PK, Ganesan A, O'Neill K, Zare H, et al. (2012) Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. Bioinformatics 28: 1009–1016. doi: 10.1093/bioinformatics/bts082

50. Aghaeepour N, Jalali A, O'Neill K, Chattopadhyay PK, Roederer M, et al. (2012) RchyOptimyx: cellular hierarchy optimization for flow cytometry. Cytometry A 81: 1022–1030. doi: 10.1002/cyto.a.22209

51. Zare H, Bashashati A, Kridel R, Aghaeepour N, Haffari G, et al. (2012) Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. Am J Clin Pathol 137: 75–85. doi: 10.1309/ajcpmmlq67yomgew

52. Qiu P (2012) Inferring phenotypic properties from single-cell characteristics. PLoS ONE 7: e37038 doi:10.1371/journal.pone.0037038.

53. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, et al. (2012) Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol 30: 858–867. doi: 10.1038/nbt.2317

54. Bashashati A, Johnson NA, Khodabakhshi AH, Whiteside MD, Zare H, et al. (2012) B cells with high side scatter parameter by flow cytometry correlate with inferior survival in diffuse large B-cell lymphoma. Am J Clin Pathol 137: 805–814. doi: 10.1309/ajcpgr8bg4jdvowr

55. Murphy RF, Chused TM (1984) A proposal for a flow cytometric data file standard. Cytometry 5: 553–555. doi: 10.1002/cyto.990050521

56. International Society for Advancement of Cytometry (2013) International Society for Advancement of Cytometry. Available: http://isac-net.org/. Accessed 5 March 2013.

57. Dean PN, Bagwell CB, Lindmo T, Murphy RF, Salzman GC (1990) Introduction to flow cytometry data file standard. Cytometry 11: 321–322. doi: 10.1002/cyto.990110302

58. Seamer LC, Bagwell CB, Barden L, Redelman D, Salzman GC, et al. (1997) Proposed new data file standard for flow cytometry, version FCS 3.0. Cytometry 28: 118–122. doi: 10.1002/(sici)1097-0320(19970601)28:2<118::aid-cyto3>3.0.co;2-b

59. Spidlen J, Moore W, Parks D, Goldberg M, Bray C, et al. (2010) Data File Standard for Flow Cytometry, version FCS 3.1. Cytometry A 77: 97–100. doi: 10.1002/cyto.a.20825

60. Leif RC, Spidlen J, Brinkman RR (2009) Cytometry Standards Continuum [7264-16]. In: SPIE BIOS Proc. Vol. 6859. 8–12 February 2009. Available: http://newportinstrument.com/cytometryml/pdf/Cytometry_Stds_Continuum08.pdf. Accessed 5 March 2013.

61. International Society for Advancement of Cytometry (2008) Analytical Cytometry Standard NetCDF Conventions for List Mode Binary Data File Component. Available: http://flowcyt.sourceforge.net/netcdf/latest.pdf. Accessed 5 March 2013.

62. Spidlen J, Shooshtari P, Kollmann TR, Brinkman RR (2011) Flow cytometry data standards. BMC Res Notes 4: 50. doi: 10.1186/1756-0500-4-50

63. International Society for Advancement of Cytometry (2012) Classification Results File Format (Report). Available: http://flowcyt.sourceforge.net/clr/latest.pdf. Accessed 5 March 2013.

64. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80. doi: 10.1186/gb-2004-5-10-r80

65. Bioconductor (2013) Bioconductor FlowCytometry view. Available: http://www.bioconductor.org/packages/release/BiocViews.html#___FlowCytometry. Accessed 11 July 2013.

66. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. (2006) GenePattern 2.0. Nat Genet 38: 500–501. doi: 10.1038/ng0506-500

67. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al.. (2012) GenePattern Flow Cytometry Suite. Available: http://www.broadinstitute.org/cancer/software/genepattern/modules/flow_cytometry. Accessed 11 July 2013.

68. Lee JA, Spidlen J, Boyce K, Cai J, Crosbie N, et al. (2008) MIFlowCyt: the minimum information about a Flow Cytometry Experiment. Cytometry A 73: 926–930. doi: 10.1002/cyto.a.20623

69. Kotecha N, Krutzik PO, Irish JM (2010) Web-based analysis and publication of flow cytometry experiments. Curr Protoc Cytom Chapter 10: Unit 10.17. doi: 10.1002/0471142956.cy1017s53

70. Spidlen J, Breuer K, Rosenberg C, Kotecha N, Brinkman RR (2012) FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. Cytometry A 81: 727–731. doi: 10.1002/cyto.a.22106

71. Spidlen J, Breuer K, Brinkman R (2012) Preparing a Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) compliant manuscript using the International Society for Advancement of Cytometry (ISAC) FCS file repository (FlowRepository.org). Curr Protoc Cytom Chapter 10: Unit 10.18. doi: 10.1002/0471142956.cy1018s61

72. Spidlen J (2013) Listing of FlowRepository public data sets. Available: http://flowrepository.org/public_experiment_representations. Accessed 5 July 2013.

73. FlowCAP - Flow Cytometry: Critical Assessment of Population Identification Methods. http://flowcap.flowsite.org. Accessed 15 March 2013.

74. Aghaeepour N (2012) IDCRP's HIV Natural History Study data set. Available: https://flowrepository.org/id/FR-FCM-ZZZK. Accessed 3 March 2013.

75. Craig FE, Brinkman RR, Eyck ST, Aghaeepour N (2013) Computational analysis optimizes the flow cytometric evaluation for lymphoma. Cytometry B Clin Cytom E-pub ahead of print. doi:10.1002/cyto.b.21115.

76. Villanova F, Di Meglio P, Inokuma M, Aghaeepour N, Perucha E, et al. (2013) Integration of lyoplate based flow cytometry and computational analysis for standardized immunological biomarker discovery. PLoS ONE 8: e65485 doi:10.1371/journal.pone.0065485.

77. Maecker HT, McCoy JP, Nussenblatt R (2012) Standardizing immunophenotyping for the Human Immunology Project. Nat Rev Immunol 12: 191–200. doi: 10.1038/nri3229

78. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647–657. doi: 10.1038/nrg2857